IEEE International Conference on Industry 4.0, Artificial Intelligence and Communications Technology (IAICT), Bali, Indonesia, July 2022.

A Survey of Machine Learning Approaches for Detecting Depression Using Smartphone Data

Zahra Solatidehkordi
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
g00059068@aus.edu

Jayroop Ramesh
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
b00057412@aus.edu

Michel Pasquier
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
mpasquier@aus.edu

Assim Sagahyroon
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
asagahyroon@aus.edu

Fadi Aloul
Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
faloul@aus.edu

***Abstract*—Depression is one of the most common mental health issues worldwide and has only become more widespread after the emergence of the Covid-19 pandemic. Although depression can be treated through various methods, it often goes undiagnosed and therefore untreated, forcing individuals to go through life with a condition that is nothing short of debilitating. With mobile phones being an integral part of people's lives, they can provide valuable information about a person's habits and behaviors, which can then be used to detect depressive tendencies. This paper provides a review of several studies conducted in recent years on the possibility of using machine learning and smartphone data to detect depression.**

***Keywords*—Depression, mental health, machine learning, smartphones**

I. Introduction

Depression is one of the leading causes of the global health-related burden [1]. According to the World Health Organization, more than 280 million people worldwide suffer from depression [2]. In 2020, the Covid-19 pandemic caused an additional 53.2 million cases of major depressive disorder globally, an increase of 27% compared to pre-Covid statistics [3]. Although depression can be treated through therapy and medication, it often goes undiagnosed due to issues such as social stigma and inaccurate assessment methods [4]. Only around 50% of cases are identified by primary care physicians [5]. This paper will discuss the possibility of using machine learning and smartphone data to detect depression in individuals. Studies have shown that there is correlation between depression and mobile phone usage characteristics; for example, people suffering from depression have been found to have fewer contacts and make fewer calls [6]. Smartphones offer a unique opportunity for mental health screening; they are closely connected to people's personal lives and can provide considerable insight on their routines, habits, activities and interactions. Additionally, smartphones are capable of collecting continuous, moment-by-moment data over long periods of time. Machine learning algorithms can be appropriate tools for detecting depression using this data as they are able to capture nonlinear and complex relationships between the dependent and independent variables [6].

One of the concerns of developing mental health apps is privacy. In a survey conducted by Lipschitz et al. about mental health assessment and monitoring apps, 59.1% of the 400 participants expressed concerns about data privacy [7]. Dogrucu et al. [8] conducted a willingness-to-disclose survey where 202 participants were asked about their willingness to share different types of data with a medical professional. The data that the participants were least likely to share were chat message contents and browser history, with more than half of the participants choosing the "slightly unwilling" or "completely unwilling" options. On the other hand, they were more willing to share voice clips, photos of themselves, GPS data and app usage data. Another concern of using smartphones for mental health assessment is the high dropout rate of applications that need active user engagement [9]. Thus, using data which are passively collected without the need for interaction from the user may be preferable.

This paper will provide a review of recent machine learning models developed for the detection of depression using smartphone data.

II. Background

The following section provides information about the machine learning models, performance metrics and mental health assessment tests mentioned in this review.

A. Machine Learning Models

The algorithms discussed in this paper are supervised learning models used for classification, in which the model is trained on a dataset containing class labels and then has to

determine the correct label of newly presented instances during testing and real-life use.

1) Support Vector Machine (SVM): SVM is a supervised machine learning algorithm in which instances are represented as points in space and classes are separated by a hyperplane [8]. It can be used for both classification and regression. Different kernel functions such as the radial basis function (RBF) can be applied to SVM to make it capable of classifying nonlinear datasets [4].

2) Ensemble Methods: Ensemble models are algorithms that create an improved model by combining several learner models. Random forest is a parallel ensemble technique, in which base learner models are generated in parallel and independently before being combined. Random forest generates multiple decision trees, with each tree using a random subset of the data and/or features for training. The predictions from the multiple decision trees are then combined [8]. Boosting is an ensemble technique where models are not independent but seek to cover previous models' weaknesses. Adaptive boosting (AdaBoost) is a boosting method that uses decision trees as learner models. In gradient boosting, the models are built sequentially, each improving the previous model, resulting in more accurate results. Extreme gradient boosting (XGBoost) is an example of gradient boosting which uses decision trees as learner models.

3) K-Nearest Neighbors: The K-nearest neighbors algorithm maps data into a multi-dimensional space and classifies each new instance based on the class labels of the K instances which are closest to it. The label chosen for the new instance is the majority label of the K neighbors [8].

B. Performance Metrics

For all the models reviewed in this paper, the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are calculated. These values are then used to calculate the following performance metrics:

1) Accuracy: The proportion of correctly classified instances to the entire dataset, calculated as seen in (1). The accuracy may not offer a complete assessment of the model's performance compared to other metrics such as the F1 or balanced accuracy as it gives the same importance to true negatives and true positives [10].

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \tag{1}$$

2) Precision or Positive Predictive Value: This metric represents the proportion of correctly classified positive instances to all positive-classified instances [10]. It showcases the ability of the model to correctly identify depressed individuals.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2}$$

1) Recall or Sensitivity or True Positive Rate: This metric represents the proportion of correctly classified positive instances to all actually positive instances [6].

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

2) Specificity or True Negative Rate: This metric represents the proportion of correctly classified negative instances to all actually negative instances; i.e., the percentage of nondepressed individuals who are correctly classified as nondepressed [6, 8].

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{4}$$

3) F1: The F1 score is the harmonic mean of the precision and recall, calculated as seen in (5). It is a popular metric within psychopathology as it's a balance between the true positive rate and positive predictive value [10]. It is suitable for unbalanced data.

$$F1 = \frac{2(precision)(recall)}{precision+recall} \tag{5}$$

4) Balanced Accuracy: This metric is calculated by taking the average of the sensitivity and specificity as described by (6). It is appropriate for datasets with class imbalances as it takes into account the correctly classified instances for both classes in equal measures [6]. In the context of this paper, the percentage of nondepressed individuals is significantly higher than depressed individuals in all datasets, as such using the balanced accuracy is appropriate.

$$\frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP+FN}} + \frac{\mathrm{TN}}{\mathrm{TN+FP}}\right) \tag{6}$$

5) False Positive Rate: Represents the proportion of incorrectly classified positive instances to all actually positive instances [10].

$$False\ Positive\ Rate = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{7}$$

6) Area under the curve (AUC): AUC is a measure of how well the classifier distinguishes between classes. The ROC curve is created by plotting the false positive rate and recall at various classification threshold values. AUC provides an aggregated measure of the performance of the model based on the ROC curve [10].

C. Psycho-metric Tests

The following tests are the measures used to label the training datasets. The results of these tests are numbers that indicate different levels of severity based on the range they fall into. However, since the machine learning models discussed in this paper only perform binary classification, each study chooses a cut-off point to separate depressed and nondepressed individuals.

1) Beck Depression Inventory Second Edition (BDI-II): A 21-item questionnaire meant to assess the existence of depressive symptoms and their severity as listed in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders [6]. Each question has a score of 0 to 3, making the total score range 0 to 63; with a total score of 0-13 indicating no depression, 14-19 indicating mild depression, 20-28 moderate and 29-63 severe.

2) Patient Health Questionnaire 8 (PHQ-8): PHQ-8 is well-established as a valid diagnostic measure of depression and is widely used in both research and clinical settings [8, 11]. It consists of 8 multiple choice questions concerning the psychological state of a person over the last 14 days. The questions involve topics such as the user's energy levels, concentration levels, appetite and quality of sleep. Each question is worth 3 points, making the minimum score of the test 0 and the maximum score 24. A total score of 0 to 4 represents no depression; 5 to 9, mild depression; 10 to 14, moderate; 15 to 19, moderately severe; and 20 to 24, severe [11].

3) Patient Health Questionnaire 9 (PHQ-9): The PHQ-9 consists of the same questions as the PHQ-8 with an additional question regarding suicide and self-harm which is used to measure suicide ideation [11]. It is the gold standard for detecting depression and measuring its severity worldwide. The score range of the PHQ-9 is 0 to 27, with the 20 to 27 range denoting severe depression. The other ranges are identical to the PHQ-8 test [8].

4) Quick Inventory of Depressive Symptomatology Questionnaire (QIDS): The QIDS questionnaire is more comprehensive than the PHQ-9 and contains 16 questions regarding depressive symptoms experienced over the last week. It covers topics similar to the PHQ-9 such as appetite, interests, energy, concentration, self-criticism and suicide ideation; however some symptoms may have more than one question dedicated to them. Each question has a score between 0 to 3 with ranges of 0 to 5 for no depression, 6 to 10 for mild depression, 11 to 15 for moderate depression, 16 to 20 for severe depression and over 21 for very severe depression [12].

III. Literature Survey

Authors Razavi, Gharipour and Gharipour [6] propose a model which uses mobile phone usage metadata to detect depression. This model does not use content attributes information such as the content of messages or social media posts, but uses routines and patterns of mobile phone usage instead. This data include average total daily usage, average number of calls received and initiated daily, average duration of calls, average number of text messages sent and received daily, number of contacts saved on the phone and average time spent on web browsing or apps. The study had 412 participants, 210 women and 202 men, all older than 18 and residents of the United States, with an average age of 40.27. The mobile usage statistics of the participants were recorded using the "Callistics" and "StayFree" apps and reported to the authors over the last 14 days before the survey. For labeling, the participants were asked to complete the BDI-II test; scores above 13 were labeled as depressed and scores below or equal to 13 were labeled as nondepressed. The model was tested using the holdout method. The dataset was split into a testing set (80% of the data) and a testing set (20% of the data). 10-fold cross validation was used on the training set to tune hyperparameters. The testing process was repeated 10 times with random training and testing partitions and the results were then averaged. A balanced random splitting method was used to ensure that the data distribution in the training and testing sets were similar. The main performance metric used to gauge the success of the model was balanced accuracy due to the class imbalance in the data; according to the BDI-II test results, 77.4% of the participants were not depressed. The classifiers tested were K-nearest neighbor, linear regression, random forest and SVM-RBF. The two most successful models were random forest with a balanced accuracy of 0.768 and BGM with a balanced accuracy of 0.723. When tested with age and gender added as attributes, the balanced accuracies of the two models increased to 0.811 and 0.766 respectively. The model with the lowest performance was linear regression with a balanced accuracy of 0.661 without age and gender and 0.699 with age and gender. The most important attributes in the performance of the random forest model were found to be average number of calls daily followed by average total daily usage and number of contacts saved on the phone. Additionally, after analyzing the misclassified samples, it was found that 81.2% percent of misclassified instances had borderline BDI-II test scores.

Opoku et al. [4] used behavioral markers from smartphone data to develop their model. The data was collected using the Carat Android app over a period of 22 days and included battery consumption data, internet connectivity, foreground app usage, and screen lock and unlock logs. A total of 22 features were extracted from this data and used in the model, all of which were calculated on the day level, such as daily lock/unlock count. The features extracted were as follows:

- Entropy: Entropy was calculated to quantify the variability, randomness and complexity in phone behavior states. For example, the screen status entropy captures the frequency and distribution in the transition of phone screen on and off states in a day.
- Regularity Index: Captures routines in the participants' behaviors by finding similarities in phone states at the same hours in all days.
- Standard Deviation and Counts: The standard deviation describes the variation in daily behavior between 4-day epochs. The count calculates the daily count of phone status, such as daily screen_on count and daily screen_off count.

The test used for labeling was the PHQ-8 test. The dataset used in this paper was significantly more unbalanced

in terms of gender compared to the previous one – out of 629 participants, 10.97% females and 86.8% males. The age of the participants ranged from 18 to over 65 years old. Based on the PHQ-8 test, 83.19% of the participants were nondepressed (score < 10) and 16.81% were depressed (score ≥ 10). The machine learning algorithms used in this paper were random forest, SVM-RBF, XGBoost, K-nearest neighbor and logistic regression. Stratified and nested cross validation was used in testing, with 3 folds in the inner cross validation and 10 folds in the outer cross validation. The class imbalance was handled using the synthetic minority over-sampling technique (SMOTE) which creates synthetic data for the minority class to balance the training set. The main performance metric used in this paper was the F1 score. The best performing model was XGBoost with an F1 score of 94% without age and gender as attributes and 95.27% with age and gender as attributes. The next best results were generated by the random forest algorithm with F1 scores of 93.41% and 95.20%. The feature importance analysis showed that the most important features for the XGBoost model were the internet regularity index, screen_on count and screen regularity index, while the most important features for the random forest model were the screen status entropy, screen regularity index and screen_off count. The results of this study showed a high correlation between screen lock/unlock patterns and depression, where individuals with depression are more likely to lock and unlock their phones routinely and randomly. Additionally, there was a strong correlation between internet usage and depression.

Tlachac and Rundensteiner [10], conducted a model on detecting depression using text message reply latency. Similarly to the previous studies, the authors took into account the privacy concerns of users and chose to collect only metadata features and not the content of the messages. The measure used for labeling was the PHQ-9 test with the threshold being set at a score of 10. The data was collected over a span of two weeks with a dataset of size 68, with all participants having replied to at least two messages within the time period. The only information used were the dates and direction of text messages (received or sent). From this data, the authors calculated the reply latency, which is the time between a 'received' and 'sent' message in seconds. For each individual, nine features were extracted from the data; the minimum and maximum latencies, the 10%, 25%, 50%, 75% and 90% quantiles of the reply latencies, the number of contacts responded to, and the number of 'sent' messages in the two-week timespan. Down-sampling was used to balance the two classes, which involves removing some of the majority class instances in the dataset. 5-fold cross validation was used for testing. The experiment procedure was repeated 100 times and the average of the results was calculated. The algorithms chosen for the experimentation were K-nearest neighbor, linear regression, SVM, random forest, XGBoost and AdaBoost. The most successful models were K-nearest neighbor with F1/AUC scores of 0.68/0.70 and XGBoost with F1/AUC scores of 0.67/0.72. The K-nearest neighbor algorithm leveraged 8 principal components while the XGBoost algorithm leveraged only the first principal component. The latter is preferred for implementation as it indicates a lower number of variables needed for the model, so XGBoost was chosen as the overall preferred method. The results of the study showed that the features and the PHQ-9 scores were correlated, with depressed individuals having a longer reply time and fewer contacts/responses.

Rafail-Evangelos et al. [9] proposed a model which uses touchscreen typing pattern analysis to detect depressive tendency. As depression has a negative impact on motor function, it likely also impacts the way users interact with their phone's touchscreen and their typing patterns. The PHQ-9 test was used for labeling. The dataset size was significantly smaller than all papers reviewed so far, with 25 participants, however the classes were more balanced (11 depressed, 14 nondepressed). An application called TypeOfMood and a custom keyboard were used to record typing data for a period of two months. This data included keystroke timing information, delete rate, number of characters typed and typing session duration. The keystroke timing features used were the hold time (time between pressing a key and releasing it) and flight time (time between releasing a key and pressing the next one). The leave-one-out cross validation was used for testing. The models tested were SVM, random forest and gradient boosting classifier. The best performing model was random forest with a mean AUC of 0.89 and sensitivity/specificity of 0.82/0.86. Similarly to the Razavi, Gharipour and Gharipour model, 75% of the incorrectly classified instances had scores close to the boundary of the PHQ-9 score. Contrarily to the previous papers, the PHQ-9 score boundary was set at 5, with scores of 0-4 indicating no depression and scores of 5-15 indicating mild to moderate depression, which means this model did not cover the 16-27 range indicating severe depression. The authors tested the model again with the entire PHQ-9 spectrum and a cutoff point of 10 which this time resulted in the gradient boosting classifier generating the best AUC score (0.81). This model only needs a total of 50 typing sessions with at least 8 keystrokes each to generate a stable result. The most important attribute was found to be the hold time. Participants with depression had longer hold times, indicating a slower motor reaction time.

Yue et al. [13] proposed a model that uses location data for depression prediction. Location data is collected from GPS and WiFi association records on smartphones through the LifeRythm app developed by the authors for both Android and IOS phones. The WiFi association records contain the address of the wireless access point (AP). Since a phone has to be close to an AP for association, the location of the AP can be used to approximate the location of the user. The GPS and WiFi data were then fused together to reduce missing data and the following features were extracted: location variance, time spent in moving, total distance, average moving speed, number of unique locations, entropy, normalized entropy and time spent at home. "Home" is defined as the location the participant usually

spends the time between 12 and 6 am. This study used a dataset of 79 college students aged 18-25 from the University of Connecticut. 25 of the participants were Android users, 6 of which were labeled as depressed and 19 as nondepressed. 54 of the participants were iPhone users, of which 13 were labeled as depressed and 41 as nondepressed. For labeling, an assessment by a clinician was done in addition to the PHQ-9 test. The PHQ-9 threshold was set at 10. SVM-RBF was the model used for classification. Leave-one-out cross validation was used for hyperparameter tuning. The Android dataset generated an F1 score of 0.67 and balanced accuracy of 0.72, while the iPhone dataset generated an F1 score of 0.77 and balanced accuracy of 0.73.

Ware et al. [12] carried out a two-phase study using location data collected from 182 college students aged 18-25 to predict individual symptoms of depression using smartphone data. Phase 1 involved data collected passively through an app on 79 participants' smartphones from October 2015 to May 2016. Phase 2 involved meta-data gained from the university's WiFi network on 103 participants from February 2017 to December 2017. The data included the results of the PHQ-9 questionnaire collected every two weeks in phase 1 and the QIDS questionnaire collected weekly in phase 2. Additionally, all participants were assessed by a clinician at the beginning of the study and students with a depression diagnosis had additional meetings on a monthly or bimonthly basis. The smartphone data contained features such as location variance, time spent in moving, total distance, time spent at home, average moving speed and number of unique locations. SVM was used with a RBF kernel to classify the presence of each depressive symptom. A recursive wrapper-based algorithm was used for feature selection and leave-one-out cross validation was used for hyperparameter tuning. The study found that behavioral depressive symptoms such as appetite, energy and sleep as well as cognitive symptoms such as interests and concentration can be detected using smartphone location data with F1 scores as high as 0.86. These results showed the correlation between location characteristics and depressive symptoms, as the lack of energy, interests and motivation caused by depression may lead to a lower tendency to move and visit different locations. The second phase of the study involved location information collected using the AP location similarly to Yue et al.'s method. The features extracted include number of significant locations visited, number of entertainment, sports and class buildings visited, average duration spent in those buildings and number of days visiting those buildings. The algorithm used was once again SVM. This phase achieved F1 scores between 0.6 and 0.7 for various symptoms, showing the potential of using WiFi meta-data to monitor the mental wellness of a large population at a low cost.

Dogrucu et al. [8] proposed a model for instantaneous depression assessment. Contrarily to all of the models reviewed so far which required the collection of data over a period of time, this model detects depression instantaneously by using a voice sample and data that is already available on the phone. However, it is also the only model that uses relatively more private data such as the content of text messages. The features used in this model consists of the following: number of contacts, call frequency, the content of text message, twitter and Instagram statistics such as number of likes, posts, followers and followed users, GPS data and audio features. All of the data is collected instantaneously from the phone using the Moodable app developed by the authors. Only the data of the past two weeks is collected. Text messages were used to calculate the text sentiment score using a python API for natural language processing. Audio features were collected from the voice sample using the openSMILE software which performs audio signal processing and extracts features such as loudness and pitch. To collect the voice sample, participants were asked to optionally record themselves saying the sentence "the quick brown fox jumps over the lazy dog". The voice sample is the only participatory data in the study. All of the data modalities in this study were optional and participants could refuse to contribute any of them. Only the PHQ-9 test was required for participants to complete the study. Among 335 total participants, all 335 contributed at least one data modality, while only 11 participants contributed all the modalities. It was found that participants with higher PHQ-9 scores were more likely to share their data; among the 11, only 2 had PHQ-9 scores below 10. Down sampling was applied to balance out the classes and the K-nearest neighbor, random forest and SVM algorithms were used for testing. Random forest showed the highest result with an F1 score of 0.766.

Choudhary et al. [14] collected smartphone data from 558 smartphone users in South Korea over an average of 10.7 days with a standard deviation of 23.7 days. Of the 558 participants, all were Android users; 286 were women, 264 were men and 18 identified as nonbinary; 474 were aged 18-25, 29 were aged 26-35, 42 were aged 36-55 and 13 were aged over 55; 487 were Korean-speaking and 71 were English-speaking. The PHQ-9 questionnaire was used to determine ground truth labels, with 495 participants showing signs of depression. The data used in this study were nonintrusive and nonidentifiable, collected passively through the Behavidence mobile app. This data involved daily behavioral patterns and were of three main types; category use per day, frequency of events per day and average time on the phone per day. 37 features were extracted in total. The models tested were random forest regression, random forest classification, multivariate adaptive regression splines, SVM and XGBoost. The most important features were mean session time in 24 hours, average session time in social apps and average session time in miscellaneous and recreational apps. The best performing classification model was random forest. The binary model with labels none or severe achieved an accuracy score of 87% while the three-class model with labels none, mild or severe achieved an accuracy of 78%. A PHQ-9 question-specific model was also developed achieving an accuracy of 78%.

TABLE I. MODEL COMPARISONS

Source	*Data Used*	*Dataset Size*	*Time period of data collection*	*Testing*	*Labeling*	*Balancing Technique*	*Best Model*	*Max. performance metric score*	*Most Important Attribute*
[6]	Phone calls, texts, internet and app usage	412	14 days	Holdout method	BDI-II	_	Random Forest	81.1% [balanced accuracy]	Avg. number of calls per day, avg. daily usage
[4]	Screen status and internet connectivity	629	22 days	Cross validation	PHQ-8	SMOTE (synthetic minority over-sampling)	XGBoost	95.27% [F1]	Internet regularity, lock & unlock
[10]	Text message reply latency	68	14 days	Cross validation	PHQ-9	Down sampling	XGBoost	72% [AUC]	_
[9]	Touchscreen typing pattern	25	60 days	Cross validation	PHQ-9	_	Only SVM tested	89% [AUC]	Hold time
[13]	Location and activity data from GPS and WiFi (combined)	79	14 days	Cross validation	PHQ-9 and clinician	_	Only SVM tested	77% [F1]	_
[12]	Location and activity data from GPS (phase 1) and WiFi (phase 2)	182	Approx. 7 months (phase 1) and 10 months (phase 2)	Holdout method	PHQ-9, QIDS and clinician	_	Only SVM tested	86% [F1]	_
[8]	Voice, texts, GPS and social media app data	335	Instantaneous	_	PHQ-9	Down sampling	Random Forest	77.1% [F1]	Voice sample
[14]	Daily behavioral patterns	558	10.7 days (average)	Bootstrapping with 15-fold cross validation	PHQ-9	Down sampling	Random Forest	87% [accuracy]	Avg. session time, avg. time on social apps, avg. time on miscellaneous and recreational apps

IV. REFLECTIONS

Table 1 shows a general comparison of all the models reviewed in this paper. All of the models discussed in this paper generated scores above 70% in the performance metrics they used. This indicates that the concept of depression screening using smartphone data has potential. Tree-based ensemble classifiers achieved better results than linear classifiers, showing that the relationship between phone usage data and depression are nonlinear. The best performing models were random forest and XGBoost. All studies explicitly mentioned the concern of data privacy and made an effort to use information that individuals are more willing to share, with the only exception being the Dogrucu et al. model which included text message content as optional data.

This topic is relatively new and requires further research. Factors that may improve results are the following:

- Collecting data over a longer period of time, as most of the models discussed in this paper only used data collected in two to three weeks. Additionally, if the period of data collection is longer than 14 days (eg. the Rafail-Evangelos et al. model), conduct new PHQ-9 tests every 14 days as opposed to only one test that covers the last 14 days.
- Using more precise tools for labeling, such as combinations of more than one psycho-metric test and clinical diagnosis of depression by a mental health professional. This is expected to improve the reliability of ground truth labels by addressing intra- and inter- personal variability across different scales.
- Conducting studies on different geographical locations and demographic characteristics, as culture can affect both mobile usage and depression. The majority of the datasets reviewed in this paper had limited variation in education level, culture and geographic location; for example, the Razavi, Gharipour and Gharipour dataset only had United States residents and the Yue et al. dataset only had college students.
- Using larger datasets, as all of the studies reviewed had datasets of size smaller than 1000, with three of them being less than 100. The latter category raises the concern of statistical significance pertaining to

the sample size considered with respect to the overall population.

- Using datasets that are more balanced and cover a wider array of ages and education levels.

Additionally, further research can focus on the modalities of federated learning [15] and integration of electronic health records [16] for well-rounded analysis fidelity. Federated learning can preserve user privacy by collaboratively training algorithms without sharing local data, which is suitable for this context. As depression is a heterogenous condition, incorporating the dimensions of medical history can improve the depression detection capabilities offered by smartphones.

In conclusion, this work serves as an initial guiding study towards highlighting the most recent developments in machine learning domain relating to the screening of depression using smartphone data. Our future work will expand on this study by discussing novel feature extraction methods and other reproducible methologies for practical implementation.

REFERENCES

[1] D. Vigo, G. Thornicroft, and R. Atun, "Estimating the true global burden of mental illness," *The Lancet Psychiatry,* vol. 3, no. 2, pp. 171-178, 2016, doi: 10.1016/S2215-0366(15)00505-2.

[2] "Depression," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] D. F. Santomauro et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet (London, England),* vol. 398, no. 10312, pp. 1700-1712, 2021, doi: 10.1016/S0140-6736(21)02143-7.

[4] A. Kennedy Opoku, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira, "Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study," (in English), *JMIR mHealth and uHealth,* vol. 9, no. 7, 2021, doi: http://dx.doi.org/10.2196/26540.

[5] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *The Lancet,* vol. 374, no. 9690, pp. 609-619, 2009, doi: https://doi.org/10.1016/S0140-6736(09)60879-5.

[6] R. Razavi, A. Gharipour, and M. Gharipour, "Depression screening using mobile phone usage metadata: a machine learning approach," *Journal of the American Medical Informatics Association : JAMIA,* vol. 27, no. 4, pp. 522-530, 2020, doi: 10.1093/jamia/ocz221.

[7] J. Lipschitz *et al.*, "Adoption of Mobile Apps for Depression and Anxiety: Cross-Sectional Survey Study on Patient Interest and Barriers to Engagement," *JMIR Ment Health,* vol. 6, no. 1, p. e11334, 2019, doi: 10.2196/11334.

[8] A. Dogrucu *et al.*, "Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data," *Smart Health,* vol. 17, 2020, doi: 10.1016/j.smhl.2020.100118.

[9] M. Rafail-Evangelos *et al.*, "Touchscreen typing pattern analysis for remote detection of the depressive tendency," (in English), *Scientific Reports (Nature Publisher Group),* vol. 9, pp. 1-12, 2019, doi: http://dx.doi.org/10.1038/s41598-019-50002-9.

[10] M. L. Tlachac and E. A. Rundensteiner, "Depression Screening from Text Message Reply Latency," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 20-24 July 2020 2020, pp. 5490-5493, doi: 10.1109/EMBC44109.2020.9175690.

[11] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of affective disorders,* vol. 114, no. 1-3, pp. 163-173, 2009.

[12] S. Ware *et al.*, "Predicting depressive symptoms using smartphone data," *Smart Health,* vol. 15, p. 100093, 2020.

[13] C. Yue *et al.*, "Fusing Location Data for Depression Prediction," *IEEE Transactions on Big Data,* vol. 7, no. 2, pp. 355-370, 2021, doi: 10.1109/TBDATA.2018.2872569.

[14] S. Choudhary, N. Thomas, J. Ellenberger, G. Srinivasan, and R. Cohen, "A Machine Learning Approach for Detecting Digital Behavioral Patterns of Depression Using Nonintrusive Smartphone Data (Complementary Path to Patient Health Questionnaire-9 Assessment): Prospective Observational Study," *JMIR Formative Research,* vol. 6, no. 5, p. e37736, 2022.

[15] S. BN and S. Abdullah, "Privacy Sensitive Speech Analysis Using Federated Learning to Assess Depression," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 6272–6276. doi: 10.1109/ICASSP43922.2022.9746827.

[16] Z. Xu et al., "Subphenotyping depression using machine learning and electronic health records," Learning Health Systems, vol. 4, no. 4, p. e10241, 2020, doi: 10.1002/lrh2.10241.